DOCUMENT RESUME

ED 175 256 FL 010 474

AUTHOR Mepham, Michael S.
TITLE L'ordinateur et l'analyse grammaticale (The Computer and Grammatical Analysis). Series B-2.
INSTITUTION Laval Univ., Quebec (Quebec). International Center for Research on Bilingualism.
PUB DATE 67
NOTE 16p.; Paper presented at the "Stage du Conseil de l'Europe 'Langues de specialite: analyse linguistique et recherche pedagogique'" (Saint-Cloud, France, November 1967)
LANGUAGE French

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Bilingualism; *Computational Linguistics; *Computer Oriented Programs; Computer Programs; Computers; Computer Science; Deep Structure; *Generative Grammar; Grammar; *Information Processing; *Language Instruction; Language Research; *Languages for Special Purposes; Sciences; Sentence Structure; Syntax; Teaching Methods; Vocabulary

ABSTRACT

This discussion of the use of computer programming in syntactic analysis covers three major points: (1) a review of basic notions in automatic grammars; (2) a description of the grammar used in a pilot project which analysed the linguistic content of methods of teaching foreign languages; and (3) proposals on the application of the same techniques to the study of scientific vocabulary. In the first section, automatic, or generative grammar, is defined as one formulated in such a way as to be applied mechanically in the construction of sentences. In this context the discussion deals with the process of analysis of existing sentences, structural description, formulation of rules, structural ambiguity, and the process for applying rules. The second section dealing with methodology covers the development of a system of analysis of texts with the help of a computer. Things to be considered in this context are lexical ambiguity, identification of rules, levels of syntactic analysis, formulation of rules, and paraphrasing of word groups, locutions, clauses, and sentences. The third section discusses techniques used in analysis of methods as these relate to the study of scientific language, with particular emphasis on the question of words and context. (AMH)

Centre international de recherche sur le bilinguisme

publication
B-2

L'ORDINATEUR ET L'ANALYSE GRAMMATICALE

MICHAEL S. MEPHAM

1967

CIRB
ICRB

L'ORDINATEUR

ET

L'ANALYSE GRAMMATICALE

Michael Mepham
Centre international de recherches sur le bilinguisme

Communication présentée au Stage du Conseil de l'Europe *Langues de spécialité: analyse linguistique et recherche pédagogique* tenu à Saint-Cloud en novembre 1967

1967

Schéma de la communication

0. Introduction

1. Les grammaires automatiques
   1.1 Définition
   1.2 Les grammaires génératives
   1.3 L'analyse syntaxique automatique
   1.4 La description structurale
   1.5 La formulation des règles
   1.6 L'ambiguïté structurale
   1.7 Le système d'analyse

2. L'analyse des méthodes
   2.1 Le projet d'analyse des méthodes
   2.2 L'ambiguïté lexicale
   2.3 L'identification des règles
   2.4 Les niveaux d'analyse syntaxique
   2.5 La formulation des règles
   2.6 Groupes et locutions
   2.7 Propositions et phrases

3. L'étude du vocabulaire scientifique
   3.1 Le traitement automatique
   3.2 Les collocations de mots
   3.3 Les collocations au niveau des groupes.
   3.4 Les collocations au niveau des locutions
   3.5 La syntaxe des textes scientifiques

L'ORDINATEUR ET L'ANALYSE GRAMMATICALE

0. Introduction

Depuis une quinzaine d'années, on a accordé beaucoup d'importance à la possibilité d'automatisation des procédés linguistiques. La traduction automatique était pendant ce temps une des préoccupations principales des chercheurs, mais d'autres projets moins ambitieux ont pu profiter du développement des ordinateurs.

Nous parlerons ici de l'automatisation dans l'analyse syntaxique sous trois chefs principaux.

D'abord, nous repasserons quelques notions de base sur les grammaires automatiques. Ensuite, nous décrirons la grammaire utilisée dans un premier projet d'analyse du contenu linguistique des méthodes d'enseignement des langues. Finalement, nous tiendrons quelques propos sur l'application des mêmes techniques à l'étude du vocabulaire général des domaines scientifiques.

1. Les grammaires automatiques

1.1 Définition

La grammaire automatique d'une langue est une grammaire formulée de façon à être appliquée machinalement. Elle doit être formelle et rigoureuse au point qu'un robot ou un automate puisse l'utiliser. C'est le cas, justement, quand il s'agit de l'ordinateur.

1.2 Les grammaires génératives

En général, une grammaire automatique sert à deux fins. D'abord, on peut fabriquer des phrases en appliquant les règles de la grammaire. Ou on peut fabriquer des structures de phrases dans lesquelles il ne reste qu'à choisir des mots de la partie du discours appropriée pour la compléter. A cause de leur capacité de composer des phrases, les grammaires automatiques sont souvent appelées des grammaires génératives. La plupart des projets de traduction automatique incorporent une grammaire générative à l'étape de production des phrases traduites.

1.3 L'analyse syntaxique automatique

Nous nous intéressons ici à un système qui a pour but, non pas de produire des phrases, mais plutôt d'analyser des phrases existantes. Une grammaire générative peut être utilisée inversement pour trouver les structures syntaxiques qui sont à la fois conformes aux règles de la grammaire et à la séquence

des mots de la phrase analysée.

1.4 La description structurale

D'après les grammairiens de l'école structuraliste, il y a pour chaque phrase une structure syntaxique sous-jacente. Cette structure consiste en une hiérarchie de syntagmes, en un réseau de dépendance, ou en d'autres choses, selon la théorie. La structure peut être représentée de diverses manières, dont une des plus commodes pour la représentation visuelle est celle des graphes arborescents. Un système de parenthèses permet de représenter la même structure sur une seule ligne.

Donc, nous parlons ici d'une grammaire qui permet de déduire les structures des phrases. Elle est composée de règles de composition des structures et de procédés d'application de ces règles. Les procédés nous trouvent les "paraphrases" ou les descriptions structurales qui sont conformes aux règles.

1.5 La formulation des règles

Les règles de la grammaire sont de différentes sortes. Par exemple, il y a des règles de dépendance et des règles

transformationnelles pour convenir aux théories de dépendance et aux théories transformationnelles. Il est plus particulièrement question ici de grammaire des structures syntagmatiques.

Pour une grammaire des structures syntagmatiques, les règles régissent la façon de regrouper les mots d'une phrase en syntagmes, et les syntagmes en d'autres syntagmes, à des niveaux plus élevés de la hiérarchie structurale. Les règles sont souvent binaires, c'est-à-dire qu'elles régissent le groupement de deux éléments à la fois. Si les règles s'appliquent à des éléments seulement sous certaines conditions de voisinage de ces éléments, la grammaire est dite "contextuelle".

1.6 L'ambiguïté structurale

L'ambiguïté structurale est inhérente à une grammaire syntagmatique. Dans la phrase "il a vu l'homme de la rue", il y a deux analyses possibles: premièrement, "l'homme de la rue" peut être un syntagme nominal; deuxièmement, "de la rue" peut être un syntagme adverbial. La formulation des règles peut aggraver ce genre d'ambiguïté, surtout si l'on se limite à des règles binaires.

1.7 Le système d'analyse

En plus des règles, une grammaire automatique possède nécessairement des procédés d'application de ces règles. Ces procédés constituent le système d'analyse qui trouve pour une phrase les descriptions structurales qui sont conformes aux règles. Le système peut éliminer une partie des ambiguïtés structurales, et limiter le nombre de paraphrases possibles d'une phrase donnée.

Le système d'analyse peut chercher les paraphrases de maintes façons: de gauche à droite, de droite à gauche, de haut en bas, de bas en haut. Il peut chercher toutes les paraphrases pour une phrase donnée, ou celles qui satisfont des critères de priorité. Par exemple, il est possible de choisir celles qui ont le moins de niveaux hiérarchiques dans leur graphe arborescent. Souvent l'idéal est de retrouver une paraphrase unique pour chaque phrase.

2. L'analyse des méthodes

2.1 Le projet d'analyse des méthodes

Nous avons entrepris des recherches sur la description du contenu des méthodes d'enseignement des langues. Mon travail couvre le développement d'un système d'analyse des

textes à l'aide de l'ordinateur. Le système comprend la codification des textes, la consultation automatique de dictionnaires et de grammaires, l'application de diverses mesures des variables textuelles, et la préparation de résumés pour décrire le contenu.

Le programme pour la séparation des mots, décrit dans la communication de monsieur Savard, fait aussi partie du système d'analyse. En font également partie, les programmes et les dictionnaires de consultation automatique pour faire l'identification des éléments morphologiques et lexicaux. Pour l'identification des éléments grammaticaux, il a fallu préparer des programmes et une grammaire pour l'analyse syntaxique automatique des phrases. Ici, il est question de la formulation et de l'utilisation de cette grammaire automatique pour nos besoins particuliers.

2.2 L'ambiguïté lexicale

D'abord, nous voulions une seule paraphrase pour chaque phrase. Pour atteindre ceci, nous avons éliminé les ambiguïtés en faisant appliquer les règles contextuelles avant les règles de structures syntagmatiques.

Les règles contextuelles ont pour but de réduire à une seule, les identifications multiples au niveau des mots. Il y a une intervention manuelle prévue pour éliminer les ambiguïtés qui persistent.

2.3 L'identification des règles

Deuxièmement, il fallait que les termes des paraphrases soient reconnaissables par les intéressés. A cette fin, nous avons attaché une étiquette, ou un "nom", à chaque règle; ainsi, chaque noeud dans le graphe arborescent porte un nom qui le caractérise comme la catégorie grammaticale caractérise les mots.

2.4 Les niveaux d'analyse syntaxique

Nous avons choisi aussi de rattacher chaque règle à un seul niveau dans la hiérarchie. Ainsi, nous pouvons parler d'une règle qui s'applique au premier niveau pour regrouper les mots. Parallèlement, nous parlons d'un élément au premier niveau syntaxique de la hiérarchie d'une phrase.

Cette formulation des règles par niveaux est faite en vue de l'utilisateur des analyses effectuées. Chaque phénomène se retrouve au même niveau dans toutes les

phrases analysées. Il en résulte que les tabulations d'éléments structuraux sont plus compréhensibles. Un autre avantage serait la facilité de comparaison des structures à l'intérieur d'une même méthode, ou entre deux méthodes.

2.5 La formulation des règles

Ce fait de rattacher chaque règle à un niveau comporte, par contre, des désavantages. La formulation de la grammaire devient moins efficace. Il faut multiplier les règles pour décrire les mêmes structures. Par exemple, il n'est plus possible d'itérer la même règle à plusieurs niveaux.

Avec un nombre arbitraire de niveaux syntaxiques, quatre dans notre cas, il nous faut comprimer les niveaux d'arborescence. La phrase qui pourrait se représenter par un graphe à sept niveaux de profondeur ne peut pas être analysée en quatre niveaux en invoquant des règles binaires. Ainsi, nous utilisons des règles qui regroupent plus que deux éléments. Ceci a l'avantage additionnel de réduire les ambiguïtés structurales.

2.6 Groupes et locutions

Les règles de la grammaire sont du genre syntagmatique

non-contextuel aux deux niveaux syntaxiques inférieurs. Elles s'appliquent de bas en haut, c'est-à-dire, à partir des mots; et de gauche à droite. Nous appelons le premier niveau, celui des groupes, et le deuxième niveau, celui des locutions. A ces deux niveaux, quand il y a deux paraphrases possibles, celle qui utilise les règles les plus longues est retenue. Ce critère de choix favorise les paraphrases qui ont le moins de noeuds arborescents au bas niveau.

2.7 Propositions et phrases

Aux deux niveaux supérieurs, ceux des propositions et de la phrase entière, la grammaire fait appel à des règles contextuelles pour séparer la phrase en propositions. Nous évitons ainsi les problèmes d'ambiguïté et de complexité structurales à ces deux niveaux.

3. L'étude du vocabulaire scientifique

3.1 Le traitement automatique

Les techniques de traitement automatique utilisées dans l'analyse des méthodes pourraient servir à l'étude du langage scientifique. L'analyse grammaticale des textes nous permettrait d'étudier plus profondément le comportement textuel

du vocabulaire, et du même coup, de connaître les caractéristiques de la syntaxe des textes scientifiques.

Mon collègue a discuté quelques techniques pour le traitement des mots. En plus, il a proposé la consultation automatique d'un dictionnaire pour repérer les syntagmes fermés qui sont connus d'avance. Cette technique s'est prouvée efficace dans l'analyse des méthodes.

3.2 Les collocations de mots

Il y a des collocations de mots qui ne sont pas des syntagmes fermés. Une collocation peut être comprise dans une structure syntagmatique du premier niveau, du deuxième niveau ou, plus rarement, d'un niveau encore plus élevé. En principe, une partie seulement des syntagmes regroupent des mots qui vont ensemble à titre de collocation. C'est surtout aux bas niveaux syntaxiques que les collocations sont importantes.

3.3 Les collocations au niveau des groupes

Il est relativement facile d'envisager le traitement du niveau des groupes pour déceler les collocations. A chaque mot du texte, nous pourrons associer automatiquement

la règle qui gouverne son inclusion dans un groupe. En plus, nous pouvons indiquer pour chaque mot, quels sont ses voisins dans le même groupe. Ces données permettent de caractériser le comportement du mot à ces deux points de vue. Par exemple, nous pouvons préparer automatiquement un résumé qui indique pour chaque mot les différentes structures de groupe dont il fait partie. Pour chaque structure de groupe différent, il est possible de résumer les différents mots que l'on a rencontrés dans ce même groupe. Ensuite, il serait possible de reviser une telle liste pour relever les contextes qui font collocation.

3.4 Les collocations au niveau des locutions

De la même façon, il est possible de repérer les collocations qui s'étendent au-delà d'un seul groupe. Il s'agit d'assigner à chaque mot toutes les données structurales de la locution qui contient le mot. Ces données sont nécessairement plus nombreuses que pour les groupes, ce qui a pour effet d'augmenter les listes de voisinage. Un mot peut faire partie d'un nombre considérable de locutions différentes, et chaque sorte de locution peut représenter un nombre élevé de combinaisons différentes de mots. Ainsi, le rendement de l'analyse syntaxique pour l'étude des collocations diminue aux

niveaux syntaxiques supérieurs.

3.5 La syntaxe des textes scientifiques

La grammaire automatique a une application évidente dans l'analyse du langage scientifique. C'est naturellement l'étude de la syntaxe des textes scientifiques. On peut envisager des concordances des structures syntagmatiques avec leurs statistiques de distribution dans les textes. Il serait possible de connaître les caractéristiques des textes scientifiques, autant pour la syntaxe que pour le vocabulaire.